

Efficient Outcome Dependent Sampling Designs for Longitudinal Binary Response Data

Jonathan S. Schildcrout¹ and Patrick J. Heagerty²

¹ Department of Biostatistics, Vanderbilt University, Nashville, TN, USA

² Department of Biostatistics, University of Washington, Seattle, WA, USA

Warehouses of longitudinal data are widely available through ongoing cohort studies, administrative databases, electronic medical records, and others. When scientific questions arise, these resources may have all necessary information to address them except for a key adjustment or even a target covariate. When examining the relationship between a relatively rare longitudinal binary response and an exposure whose ascertainment costs are high, researchers are often interested in sampling designs that can maximize estimation efficiency while maintaining a reasonable budget. Outcome dependent sampling designs target key features of the response distribution for the purpose of addressing scientific questions efficiently. With a univariate response, case control designs permit valid and efficient parameter estimation from a prospective logistic model without acknowledging the retrospective sampling scheme. This is not possible with correlated and specifically longitudinal response data. In this presentation we will discuss a class of outcome dependent sampling designs that sample based on the sum of elements in the binary response series. We will use conditional maximum likelihood, and will show that estimation efficiency can be very high for time-varying and time-invariant covariates parameters even when only sampling a fraction of subjects from the original cohort.